
Article:
Laud, P.J. orcid.org/0000-0002-3766-7090 (2017) Equal-tailed confidence intervals for comparison of rates. Pharmaceutical Statistics. ISSN 1539-1604

https://doi.org/10.1002/pst.1813

Pharmaceutical Statistics

Equal-tailed confidence intervals for comparison of rates

Peter J. Laud*

Several methods are available for generating confidence intervals for rate difference, rate ratio, or odds ratio, when comparing two independent binomial proportions or Poisson (exposure-adjusted) incidence rates. Most methods have some degree of systematic bias in one-sided coverage, so that a nominal 95% two-sided interval cannot be assumed to have tail probabilities of 2.5% at each end, and any associated hypothesis test is at risk of inflated type I error rate. Skewness-corrected asymptotic score methods have been shown to have superior equal-tailed coverage properties for the binomial case. This paper completes this class of methods by introducing novel skewness corrections for the Poisson case and for odds ratio, with and without stratification. Graphical methods are used to compare the performance of these intervals against selected alternatives. The skewness-corrected methods perform favourably in all situations - including those with small sample sizes or rare events - and the skewness correction should be considered essential for analysis of rate ratios. The stratified method is found to have excellent coverage properties for a fixed effects analysis. In addition, another new stratified score method is proposed, based on the *t*-distribution, which is suitable for use in either a fixed effects or random effects analysis. By using a novel weighting scheme, this approach improves on conventional and modern meta-analysis methods that rely on weights based on crude estimation of stratum variances. In summary, this paper describes methods that are found to be robust for a wide range of applications in the analysis of rates.

Keywords: Confidence interval, binomial proportions, Poisson rates, difference or ratio, stratified, meta-analysis

1. INTRODUCTION

The comparison of two treatment groups with respect to a binomial proportion is commonly made in the field of medical statistics, such as in the analysis of efficacy or safety endpoints in phase III clinical trials, either to address the primary study objective, or for exploratory analysis of subgroups. Occasionally a Poisson event rate is used that is adjusted for the duration of exposure to risk, such as the number of adverse events observed per person-year exposure to treatment. Any of three comparative parameters may be used: rate difference ('RD'), rate ratio ('RR', also known as relative risk) or odds ratio ('OR', which is applicable only to binomial proportions).

In any of these situations, it is usual to perform a hypothesis test for a difference between treatments. In addition, it is important to estimate the magnitude of the treatment effect, along with a confidence interval to represent the uncertainty around the estimate, in order to consider the clinical relevance of any differences supported by the data. Sometimes, a one-sided significance test may be required against a non-zero null hypothesis (for example, in a 'non-inferiority' analysis, designed to demonstrate that the difference between a new treatment and an established one is not greater than a pre-specified margin), and this test is directly related to the location of a one-sided confidence limit. A two-sided confidence interval may be similarly employed to apply a formal test of bioequivalence.

Because of the discrete nature of the data in these types of analysis, it is not possible for a confidence interval to achieve precisely the desired coverage in all situations. In other words, a nominal '95%' confidence interval will not contain the 'true' treatment effect exactly 95% of the time, and nor will it be guaranteed to have evenly-distributed tail probabilities of 2.5% at each end. A large number of methods have been developed that aim to optimise the first of these desired coverage properties, but relatively little attention has been paid to the second, 'equal-tailed' criterion. This criterion deserves greater prominence, not least in situations where there is likely to be more interest in only one side of a confidence interval, such as when using the rate difference for adverse event rates, which can be interpreted as attributable risk. Therefore the one-sided non-coverage probabilities are of particular interest in this paper.

Miettinen and Nurminen's asymptotic score confidence limits[1] for the comparisons of binomial or Poisson rates were derived by defining, for each case, a contrast-based chi-squared test statistic (or 'score') as a function of the comparative parameter θ, and 'inverting' the statistic by setting it equal to a quantile of the chi-squared distribution

*Correspondence to: Peter J. Laud, Statistical Services Unit, University of Sheffield, Hicks Building, Hounsfield Road, Sheffield, South Yorkshire, S3 7RH, UK. Email: p.j.laud@sheffield.ac.uk

and solving for θ. For the binomial RD case, these methods have been found to have very good coverage properties.[2,3,4]

Gart and Nam[5,6] derived similar intervals for binomial RD and RR, using an 'efficient score' statistic instead of the simple contrast-based score. They noted that tail probabilities can be distributed quite unevenly, and incorporated a correction for skewness, concluding that the need for this correction was greatest in the case of RR. However, the Gart-Nam formulae can occasionally be affected by indeterminate $(0 \div 0)$ values in the score statistic, causing one of the confidence limits to be incalculable.

Laud and Dane[2] redefined the skewness correction for RD using Miettinen and Nurminen's contrast-based test statistic, which on the whole avoids indeterminate scores. The skewness correction was observed to have an appreciable effect on one-sided coverage properties for RD when sample sizes are unequal.

The purpose of this paper is to describe novel skewness-corrected methods for Poisson RD and RR, binomial OR and the single binomial or Poisson rate p. Graphical methods are used to compare the performance of this whole class of methods against other recently developed methods that have not been extensively evaluated. Then, having also applied the skewness correction for the stratified case, a further aim is to begin to explore the performance of score methods for stratified datasets, such as may be used in a meta-analysis.

The statistical properties of the unstratified methods are evaluated using the graphical methods of Laud and Dane,[2] with an emphasis on one-sided non-coverage probability. This is because the correction for skewness is designed to improve one-sided coverage in order to achieve a 'symmetrical', 'equal-tailed' or 'centrally located' interval, in other words one for which non-coverage probabilities are as close as possible to the nominal significance level $\alpha/2$ on each side. Interval location was discussed for the one-sample case by Newcombe,[7] who pointed out the importance of considering this aspect of performance in the evaluation of confidence intervals for proportions and related quantities. In the context of a non-inferiority hypothesis test, one-sided coverage is directly related to the one-sided type I error rate (hence the terms 'type I error' and 'non-coverage' are used interchangeably here).

Selection of the 'best' method is based on a 'proximate' coverage criterion (i.e. aligning mean coverage with the nominal confidence level). The alternative would be to take a strictly conservative approach (aligning minimum coverage with the nominal level), and the choice between these standpoints is a somewhat polarising issue. Reasons for choosing the proximate criterion can include concerns over 'exact' intervals being over-conservative, as well as practical considerations, but perhaps most persuasive is the argument that in most other areas of statistical practice, control of type I error rate is only achieved in an approximate sense based on the assumptions underlying the chosen statistical model.[8,9,10] Note that proximate *two-sided* coverage may be taken for granted for methods that consistently achieve proximate one-sided coverage. Therefore two-sided coverage is generally not discussed in this paper. Interval width is also not studied in detail, because interval location is considered to take precedence.

The structure of the paper is laid out as follows: Section 2 defines the general form of the skewness-corrected asymptotic score method, with the specific details of the calculations for each parameter contained in Appendix A. Some alternative methods for comparison are described, and some illustrative examples are discussed. Section 3 presents a graphical evaluation of the coverage properties of the proposed class of methods, using closed-form calculation of non-coverage probabilities. Section 4 introduces some new intervals for stratified datasets, and the performance of a number of stratified methods is evaluated using a simulation study. Overall conclusions are in Section 5. Additional details are available in the online supporting information, including an extended graphical evaluation with a range of sample sizes; the application of 'continuity corrections' for those who prefer the strictly conservative approach; and methods for identifying, quantifying and accounting for stratum heterogeneity.

2. DEFINITIONS

2.1. Score methods

Miettinen and Nurminen's confidence limits (denoted '**MN**') employed a chi-squared test statistic with one degree of freedom, which may naturally be square-rooted to be expressed equivalently as a normal test statistic (as used in the Farrington-Manning test[11]):

$$Z(\theta) = \frac{S(\theta)}{\widetilde{V}^{1/2}} \tag{1}$$

The $100(1-\alpha)\%$ confidence limits are the two values of θ satisfying $Z(\theta) = \pm z$, where z is the relevant percentile of the standard normal distribution, i.e. $z_{1-\alpha/2}$.

The specific details of the components of this function depend on the situation, as defined more fully in Appendix A. In general terms:

- θ represents the comparative parameter for RD, RR or OR (i.e. $p_1 - p_2$, p_1/p_2 or $p_1(1-p_2)/(p_2(1-p_1))$ respectively, where p_1 and p_2 are the underlying event rates in the two groups);
- $S(\theta)$ is a contrast function involving the sample observed rates $\widehat{p}_1 = X_1/n_1$ and $\widehat{p}_2 = X_2/n_2$, where the denominator $n_i, i = 1, 2$, represents a number of subjects for the binomial case, or a measure of total exposure duration in the Poisson case;

- $\widetilde{V}$ is the estimated variance of $S(\theta)$ using $\widetilde{p}_1$ and $\widetilde{p}_2$, the maximum likelihood estimates (MLEs) of the two event rates under the restriction that $\widetilde{p}_1 - \widetilde{p}_2 = \theta$ (for RD), $\widetilde{p}_1 = \widetilde{p}_2\theta$ (for RR), or $\widetilde{p}_1 = \widetilde{p}_2\theta(1 + \widetilde{p}_2(\theta - 1))$ (for OR).

Gart and Nam introduced skewness corrections for binomial RR[5] and RD,[6] originating from Bartlett[12] and Cornish and Fisher,[13] resulting in a test statistic that takes the form:

$$Z(\theta) = \frac{S(\theta)}{\widetilde{V}^{1/2}} - \frac{(z^2 - 1)\widetilde{\mu}_3}{6\widetilde{V}^{3/2}} \quad (2)$$

This function involves the estimated skewness $\widetilde{\mu}_3/\widetilde{V}^{3/2}$, where $\widetilde{\mu}_3$ is the estimated third central moment of $S(\theta)$, again based on $\widetilde{p}_1$ and $\widetilde{p}_2$. Note that Gart and Nam used a different form of $S(\theta)$, but their method of deriving the skewness correction is followed through in Appendix A for all comparisons, using the score functions given by Miettinen and Nurminen. This yields a comprehensive class of 'Skewness-Corrected Asymptotic Score' methods (denoted '**SCAS**'), which includes novel confidence intervals for OR; Poisson RD and RR; and the single binomial or Poisson rate p.

The variance and skewness estimators are both essentially functions of the comparative parameter θ, so Equations (1) and (2) resolve to functions of θ, which in most cases are nonlinear. The solutions are most easily found by iteration over θ, for example using the secant or bisection root-finding method. Further details are given in Appendix S1 (available online as Supporting Information).

An option to include a further 'continuity correction' for more conservative coverage is also possible. Details may be found in Appendices S2 and S3.4 (available online).

Equation (2) may be re-stated as a quadratic equation in $Z(\theta)$, which results in the same confidence limits for any given confidence level, but also allows the score (and hence p-value) to be calculated for a hypothesis test against any null hypothesis value θ_0 for θ, by solving for $Z(\theta_0)$ (see Appendix A.5 for details):

$$Z(\theta_0) = \frac{S(\theta_0)}{\widetilde{V}^{1/2}} - \frac{(Z(\theta_0)^2 - 1)\widetilde{\mu}_3}{6\widetilde{V}^{3/2}} \quad (3)$$

The use of a Z statistic here facilitates one-sided (e.g. non-inferiority) tests. The squared score $Z(\theta_0)^2$ provides a generalised version of the usual Pearson chi-squared test, incorporating corrections for both skewness and variance bias.

2.2. Approximate Bayesian MOVER and other methods

An alternative method for consideration is the 'Method of Variance Estimates Recovery' ('**MOVER**'),[14,15] which constructs an interval for θ from separate intervals for the individual group rates. This approach was first proposed for binomial RD by Newcombe,[4] using the Wilson score interval, which has since been shown to have a systematic bias in one-sided coverage.[16] If the equal-tailed Jeffreys method[9,17] is used instead (hence denoted '**MOVER-J**'), it might be expected to result in a more equal-tailed interval for θ. This is not an entirely novel development, but this class of methods (covering all comparisons of binomial or Poisson rates) is described together for the first time here, in order to evaluate their performance against the corresponding asymptotic score methods. In fact, **MOVER-J** can be viewed as a special case of a more general Bayesian method ('**MOVER-B**'), which allows the use of informative priors. This method can be applied to all of the rate comparisons as follows:

First, separate confidence intervals (l_i, u_i) and point estimates $\widehat{p}_i$ are generated for $p_i, i = 1, 2$, from the $\alpha/2$ and $(1 - \alpha/2)$ quantiles and median of the $Beta(X_i + a_i, n_i - X_i + b_i)$ distribution for binomial rates, or the $Gamma(X_i + a_i, 1/n_i)$ distribution for Poisson rates (where $1/n_i$ represents the scale parameter). The **MOVER-J** method applies the non-informative Jeffreys prior using $a_i = b_i = 0.5$. (Boundary modifications at $X_i = 0$ or n_i, as recommended by Brown et al,[9] are omitted here to avoid complicating the more general **MOVER-B** method.)

Then, for RD,[4] the confidence limits (L, U) for θ are:

$$L = \widehat{p}_1 - \widehat{p}_2 - \sqrt{(\widehat{p}_1 - l_1)^2 + (u_2 - \widehat{p}_2)^2},$$

$$U = \widehat{p}_1 - \widehat{p}_2 + \sqrt{(u_1 - \widehat{p}_1)^2 + (\widehat{p}_2 - l_2)^2}$$

For RR,[14]

$$L = \frac{\widehat{p}_1\widehat{p}_2 - \sqrt{(\widehat{p}_1\widehat{p}_2)^2 - l_1u_2(2\widehat{p}_1 - l_1)(2\widehat{p}_2 - u_2)}}{u_2(2\widehat{p}_2 - u_2)},$$

$$U = \frac{\widehat{p}_1\widehat{p}_2 + \sqrt{(\widehat{p}_1\widehat{p}_2)^2 - u_1l_2(2\widehat{p}_1 - u_1)(2\widehat{p}_2 - l_2)}}{l_2(2\widehat{p}_2 - l_2)}$$

The same formulae are used for OR,[18] but with each $\widehat{p}_i$, l_i and u_i replaced with $\widehat{p}_i/(1 - \widehat{p}_i)$, $l_i/(1 - l_i)$ and $u_i/(1 - u_i)$ respectively.

By adapting the parameters a_i and b_i in the derivation of (l_i, u_i), **MOVER-B** may be used to incorporate prior beliefs about p_1 and p_2 in order to construct an approximate Bayesian credible interval for θ. Other (computationally intensive) 'exact Bayesian' methods have been proposed,[19,20] which have recently been incorporated into StatXact software (version 11). These are discussed briefly in Section 3.6.

For reference, the following evaluation also considers the simple 'approximate normal' ('**AN**') methods available for each case (sometimes referred to as 'Wald' methods). Definitions of these methods, based on normal approximations for θ or $ln(\theta)$, can be found in Rothman and Greenland.[21]

2.3. Example confidence intervals

A number of real data examples are included in Appendix B for software validation purposes and to illustrate the magnitude of differences that can be observed between methods. However, it must be noted that no general statements can be made about the relative widths or locations of different methods based on any single example, because the direction of apparent differences between any two methods can fluctuate across the parameter space. To emphasise this point, it may be noted from the examples in Appendix B that the different methods are often shifted relative to each other, and this shift is not always in the same direction. The following evaluation is designed to draw out the underlying trends in these location shifts, in order to identify which method has the 'correct' location properties 'on average'.

Furthermore, overinterpretation of the relative widths of such example intervals should be avoided, not least because the narrowest interval is not necessarily the best one, as discussed in Section 5. Reductions in interval width are often achieved at the expense of a less central interval location. Also, rather than surmising that one interval is 'too wide', the possibility should be considered that the other intervals are 'too narrow'. In other words, a wider interval may be an accurate reflection of the lack of information contained in a given dataset.

3. EVALUATION IN THE SINGLE-STRATUM CASE

The performance of the skewness-corrected asymptotic score methods is evaluated below, using the methods of calculation and graphical display described in Laud and Dane.[2] Briefly, for each parameter space point (PSP) (p_1, p_2), one-sided non-coverage probability is calculated precisely, as a sum of bivariate binomial (or Poisson) probabilities. In other words, the non-coverage rates are obtained from essentially closed-form calculations, rather than by simulation. Full details of these calculations are provided in Appendix S3 (available online). These probabilities may naturally be considered in terms of either right- or left-sided non-coverage; this article (arbitrarily) uses right-sided non-coverage probability ('RNCP') throughout, so that results correspond to the one-sided type I error rate for a non-inferiority test when the outcome variable is an undesirable event, such as mortality or adverse event rate.

Smoothed (moving average) and unsmoothed representations of the 3-dimensional probability surface are plotted in two dimensions using a colour-shaded contour plot, clearly showing the proximity of RNCP to the nominal $\alpha/2$ over a range of the parameter space. In the Poisson case, the theoretical parameter space extends to $+\infty$ on both axes, but only the $[0, 1]$ interval is shown. Good equal-tailed performance is indicated by regions of orange colour, corresponding to RNCP being within $[0.9\alpha/2, 1.1\alpha/2]$. The proportion of the displayed parameter space having smoothed RNCP within this range is defined as the 'moving average % proximate', which provides a simple numeric summary measure of interval location. All methods can have some regions of the parameter space with proximate coverage, so any method could be acceptable for the 'right' specific null hypothesis. Proximate coverage across 100% of the parameter space would allow a method to be employed universally, both for general descriptive use and for hypothesis testing. This definition of 'proximate' allows error rates to be slightly above the nominal $\alpha/2$, indicated by dark orange, but predominance of pale orange colour would be preferred. Shades of red to black colour show regions where the actual type I error rate would be inflated by an unacceptable amount.

The coverage probability plots for the **SCAS** methods are displayed in Figures 1 to 3 alongside those for the **MN**, **MOVER-J**, and **AN** methods. Moderately large sample examples are given, using $\alpha = 0.05$, $n_1 = 150$ and $n_2 = 50$ in each case, demonstrating the efficacy of the skewness corrections here. Appendix S3.1 (available online) contains coverage probability plots for a variety of different sample sizes (including $n_1 = 50$ and $n_2 = 150$, which may be used to infer left-sided non-coverage rates for comparison with the RNCP presented in this section). The performance of the methods under the selected sample size conditions are summarised together in Table 1 using moving average % proximate. By this measure, **SCAS** is seen to be either the best or close to the best in every case. Similar patterns were observed for $\alpha = 0.01$ and $\alpha = 0.1$ (not shown). Similar patterns were also observed for smaller samples, except that where methods diverged from nominal coverage, they did so to a greater extent, while **SCAS** generally remained the most proximate method. There can be small areas of the parameter space where non-coverage for **SCAS** deviates from the nominal value, but these deviations are generally on the conservative side. In general, the plots speak for themselves, but some brief explanatory comments are provided in the following sections.

3.1. Rate Difference

RNCP surface plots for binomial and Poisson RD are shown in Figure 1. For a full evaluation of the performance of confidence intervals for binomial RD, see Laud and Dane, where the **SCAS** method was denoted '**GNbc**', and **MOVER-J** was referred to as '**NJ**'. In summary, **SCAS** stands out as being the most consistent method in terms of one-sided coverage, while other methods fail to achieve equal-tailed coverage even with large sample sizes.

For Poisson RD (Figure 1b), similar results were observed. In the example shown, the **AN** method has very conservative RNCP, but the associated left-sided non-coverage would be greatly inflated, and the opposite would be seen if the

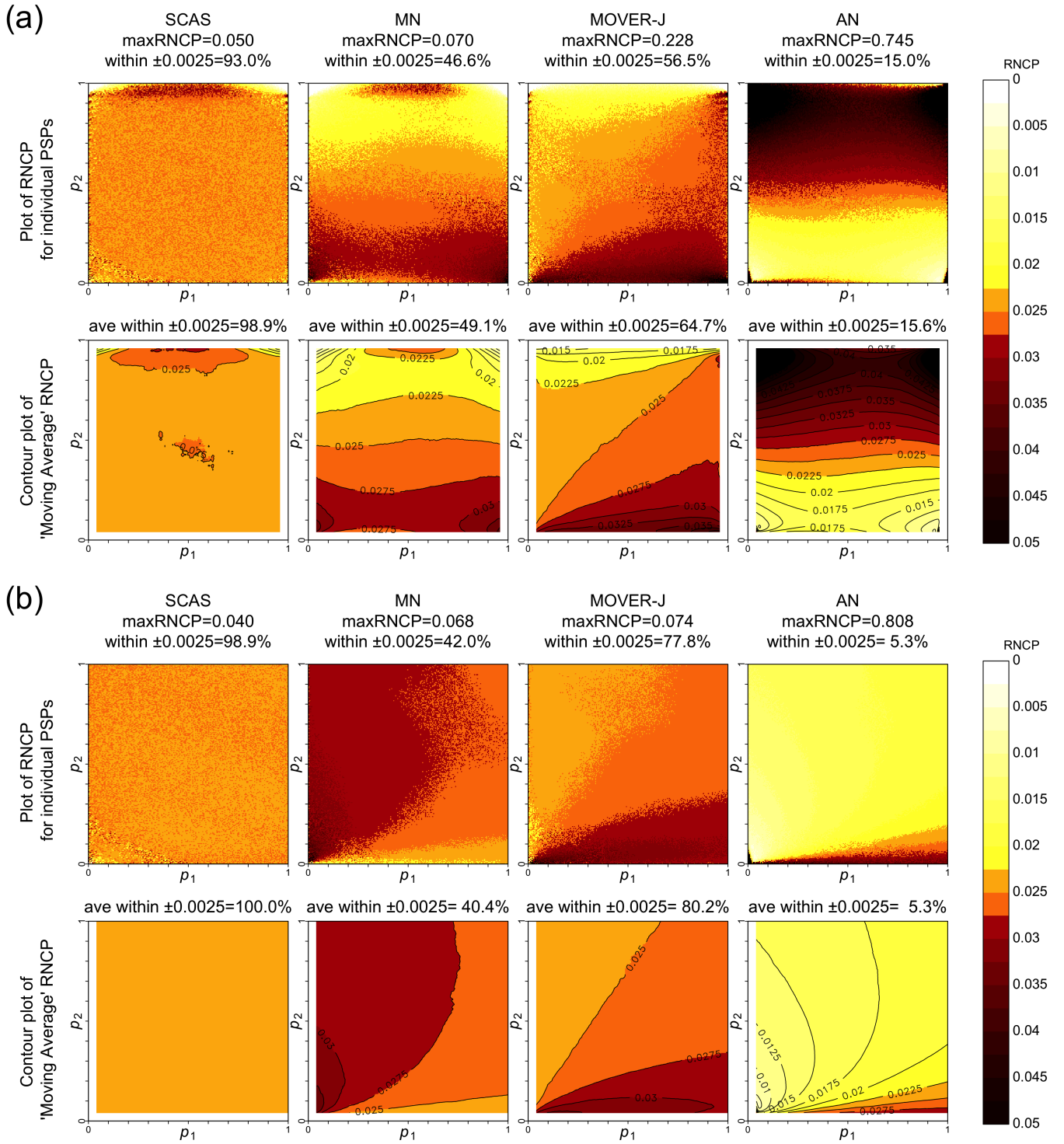

Figure 1. Rate Difference: Contour plots of right-sided non-coverage probability (RNCP) and moving average RNCP, for **SCAS**, **MN**, **MOVER-J** and **AN**, with $n_1 = 150$, $n_2 = 50$. (a) Binomial rate difference, (b) Poisson rate difference.

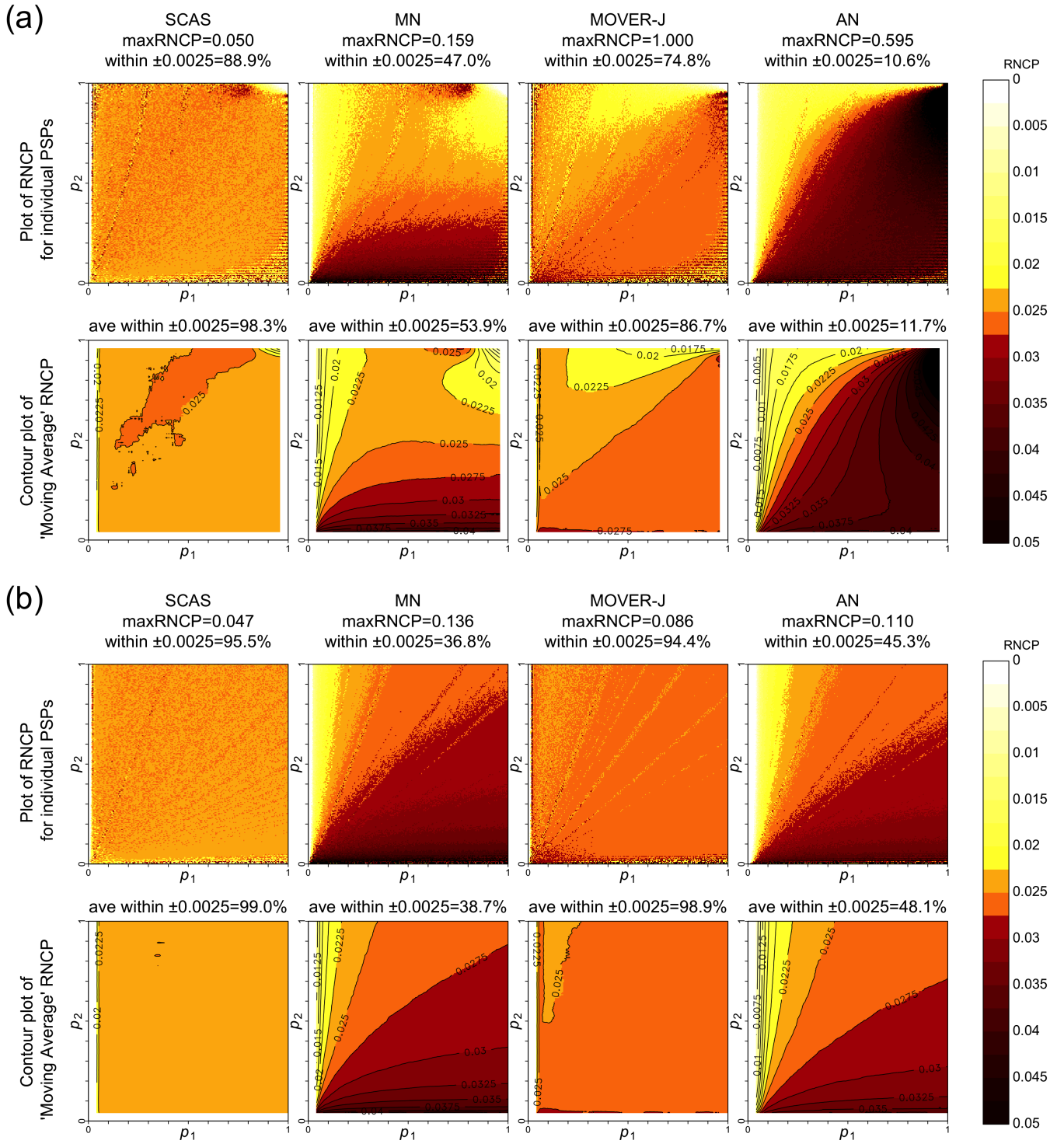

Figure 2. Rate Ratio: Contour plots of right-sided non-coverage probability (RNCP) and moving average RNCP, for **SCAS**, **MN**, **MOVER-J** and **AN**, with $n_1 = 150$, $n_2 = 50$. (a) Binomial rate ratio, (b) Poisson rate ratio.

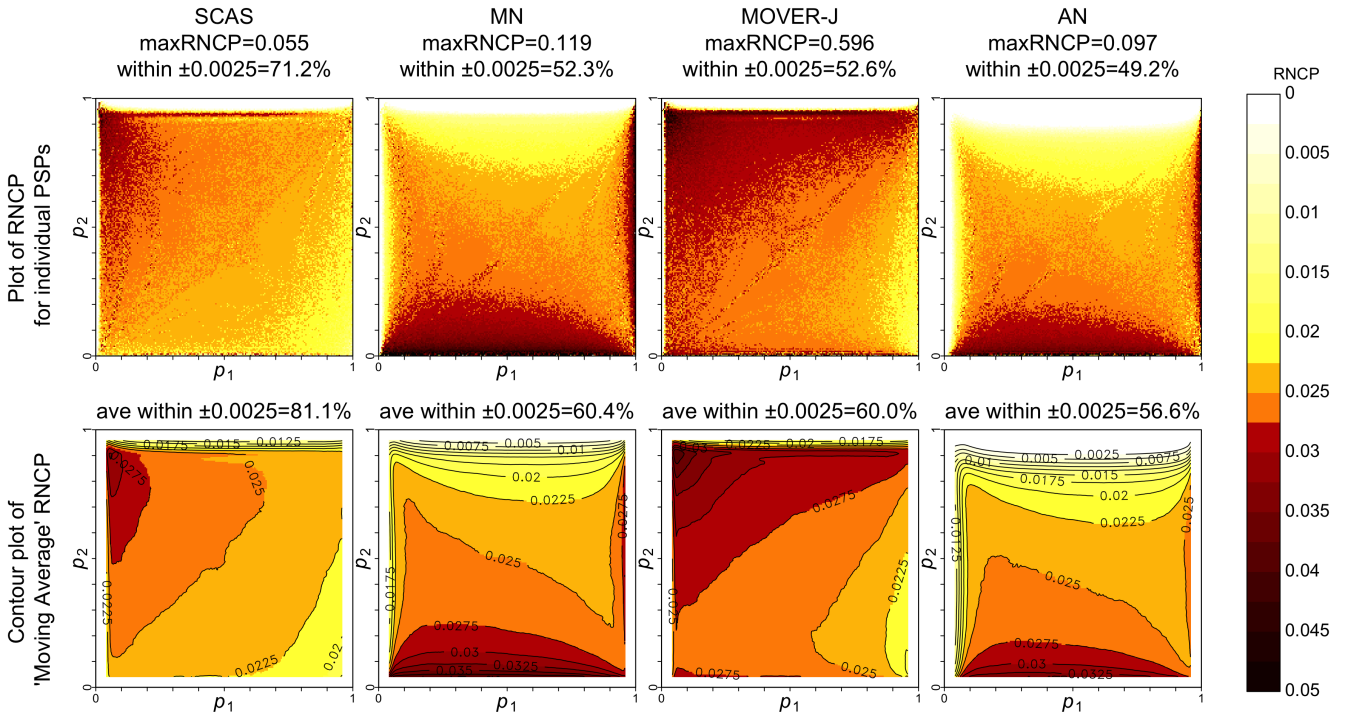

Figure 3. Odds Ratio: Contour plots of right-sided non-coverage probability (RNCP) and moving average RNCP, for **SCAS**, **MN**, **MOVER-J** and **AN**, with $n_1 = 150$, $n_2 = 50$. Odds ratio.

treatment group sizes were reversed (as shown in Appendix S3.1, available online).

3.2. Rate Ratio

With RR as the comparative parameter (Figure 2), the one-sided performance of the **MN** method is notably worse than it is for RD, reflecting the observations made by Gart and Nam.[5] Both **MN** and **AN** should be avoided here. The **SCAS** method successfully equalises the tail probabilities for both the binomial and Poisson rate ratio. The **MOVER-J** methods also perform very well for RR, although on average RNCP tends to be slightly above the nominal $\alpha/2$, and performance deteriorates with smaller sample sizes (see Table 1).

3.3. Odds Ratio

For OR (Figure 3), the proposed skewness correction unfortunately does not appear to achieve quite the same degree of consistent one-sided coverage as it does for RD and RR, but it is still superior to all of the other methods, especially when p_1 and p_2 are small. It appears that the skewness may be over-corrected when there is a large difference between p_1 and p_2.

MOVER-J also performs relatively poorly for OR. **AN** is better here than with other contrasts, although it gives an uninformative interval when there are any zero cell counts.

3.4. Rare events

For the analysis of rare events, it is the extreme lower left corner of the plots that is of interest. Appendix S3.2 (available online) contains a selection of expanded surface plots showing RNCP for the region with p_1 and p_2 less than 0.05. Plots of the left-sided non-coverage probabilities (LNCP) are also shown here, since they cannot be inferred from the RNCP plots as they can elsewhere. In general, it can be seen that the **SCAS** method maintains the most consistent performance in this region, whereas the other methods all have large regions of inflated non-coverage rates. In particular, the **SCAS** method for OR performs well in this region, despite imperfections elsewhere in the parameter space.

3.5. Single binomial or Poisson rate

The coverage properties of the **SCAS** method for the single binomial or Poisson rate are compared against the **Jeffreys**, **Score** and **Wald** intervals (see Cai[16] for definitions) in Appendix S3.5 (available online).

The **Score** and **Wald** intervals both have a systematic bias in one-sided coverage, which is corrected by the other two methods. The improvement achieved by **SCAS** over the

Table 1. Summary of one-sided moving average % proximate for various sample sizes

	Binomial RD					Poisson RD				
	30,30	45,15	100,100	150,50	50,150	30,30	45,15	100,100	150,50	50,150
SCAS	98.1	86.0	100.0	98.9	99.0	99.2	99.4	100.0	100.0	99.6
MN	64.5	25.6	85.3	49.1	49.1	61.2	13.6	91.5	40.4	26.7
MOVER-J	28.3	36.2	51.8	64.7	64.8	62.8	66.0	95.4	80.2	85.0
AN	18.6	8.9	45.7	15.6	15.6	29.5	3.1	48.2	5.3	5.6
	Binomial RR					Poisson RR				
	30,30	45,15	100,100	150,50	50,150	30,30	45,15	100,100	150,50	50,150
SCAS	89.5	87.0	97.8	98.3	95.1	91.9	93.3	97.9	99.0	95.4
MN	44.6	32.8	63.0	53.9	53.9	40.5	16.1	60.1	38.7	33.9
MOVER-J	47.5	39.2	95.4	86.7	69.9	88.9	86.8	98.9	98.9	95.8
AN	8.1	5.9	16.8	11.7	10.9	34.1	28.4	58.2	48.1	21.7
	OR									
	30,30	45,15	100,100	150,50	50,150					
SCAS	51.9	49.7	81.5	81.1	81.3					
MN	60.8	40.8	76.0	60.4	60.3					
MOVER-J	45.0	29.1	71.1	60.0	76.1					
AN	51.4	32.1	72.5	56.6	56.6					

Jeffreys method is small but significant: both methods have very consistent moving average RNCP over the whole range of θ, but it tends to be slightly higher than the nominal level for the Jeffreys interval. The **SCAS** method may be utilised for a transformed score interval for paired odds ratio, as mentioned in the discussion. It may also be of use if a stratified interval for a single rate is required, by applying the formula described in Section 4.

3.6. MOVER-J compared with other Bayesian methods

In addition to **MOVER-J**, a number of alternative Bayesian methods are available, and the following observations about them may be of interest.

Firstly, the so-called 'exact Bayesian' intervals[20] are not 'exact' in the usual sense regarding achieving a minimum coverage criterion, but rather in the sense that the tail probabilities, given a particular prior distribution, are calculated exactly. The default settings in StatXact use uniform (rectangular) priors for the p_is (i.e. $a_i = b_i = 1$), as suggested by Nurminen and Mutanen.[20] Agresti and Min[22] found that the diffuse Jeffreys prior ($a_i = b_i = 0.5$) achieved much better frequentist performance, but they also noted that even with the Jeffreys prior, the Bayesian methods were inferior to score-based methods.

For the special case of Poisson RR, it is possible to obtain a Bayesian confidence interval by transforming a Jeffreys interval for a single *binomial* proportion, conditioning on the total number of events X, as described by Barker and Cadwell[23] (although their formulae need a slight modification if the exposure times n_1 and n_2 are unequal).

Figures in Appendix S3.3 (available online) demonstrate that the **MOVER-J** methods have very similar coverage properties to the above Bayesian methods where Jeffreys priors are used, implying that **MOVER-B** should be an adequate substitute for the exact Bayesian methods with informative priors. The figures also show that one-sided coverage is quite poor when uniform priors are used.

In summary, for a Bayesian analysis using a non-informative prior, the diffuse Jeffreys priors ($a_i = b_i = 0.5$) are recommended for all contrasts for both the binomial and Poisson case, and the **MOVER-J** methods are simple and adequate for this purpose. These methods are outperformed by the **SCAS** method in terms of frequentist coverage probabilities, but **MOVER-B** allows the facility to incorporate informative priors for the event rates.

4. STRATIFIED METHODS

In order to adjust for a stratification factor that is expected to affect the underlying true overall event rate $\bar{p} = (p_1 + p_2)/2$, a confidence interval is constructed using a weighted average of stratum-specific scores. A number of different weighting schemes are available, two popular choices being Mantel-Haenszel ('MH') weights (which essentially reflect the sample size in each stratum), and inverse variance ('IV') weights. For the asymptotic score methods, a new weighting strategy is possible, using the inverse variance of the score ('IVS') to define weights as a function of θ, by inverting the MLE of the stratum variances $\widetilde{V}_j$. Other options include the hybrid 'Minimum Risk' ('MR') weights described by Mehrotra and Railkar[24]; inverse variance weights using some other modified variance estimate using shrinkage towards a pooled estimate[25]; or some iteratively calculated weights as given by Miettinen and Nurminen.[1]

See Appendix S4.1 (available online) for further discussion of weighting schemes, including some important notes on the MH weights for the score method, which are not always the same as the conventional MH weights.

With stratum weights w_j, and $W = \sum_j w_j$, a stratified **SCAS** confidence interval is found by solving:

$$Z(\theta) = \sum_j \left[\frac{(w_j/W)S_j(\theta)}{\widetilde{V}_\cdot^{1/2}} - \frac{(w_j/W)^3 \widetilde{\mu}_{3j}(z^2-1)}{6\widetilde{V}_\cdot^{3/2}} \right] = \pm z \quad (4)$$

where $\widetilde{V}_\cdot = \sum_j (w_j/W)^2 \widetilde{V}_j$; and $S_j(\theta)$, $\widetilde{V}_j$ and $\widetilde{\mu}_{3j}$ are the stratum-specific quantities using the definitions in Appendix A.

The two solutions of this equation are found by iteration over θ as before. Note that, as with the single stratum case, the same formula can again be used to apply a significance test against any null hypothesis value of the overall treatment effect θ, by replacing z with $Z(\theta)$ and solving the resulting quadratic.

Unlike in other meta-analysis methods, no adjustment is generally necessary for zero cell counts here, even for RR and OR, although they do pose a problem if there are any 'double-zero' strata, containing no events in both arms. See Appendix S1.3 for further details.

4.1. Heterogeneity/interaction tests and 'Random effects' intervals

The calculation of 'fixed effects' stratified confidence intervals described above relies on the assumption that θ is constant across strata (i.e. homogeneous). In other words, the null hypothesis under consideration is that $\theta_j = \theta_0$ for all j. Methods for testing this assumption, and for visualisation and quantification of heterogeneity (i.e. treatment-by-stratum interactions) are described in Appendix S4.2 (available online).

It may be desirable to incorporate any stratum heterogeneity into the calculation of the confidence interval, using a 'random effects' meta-analysis approach in order to present an interval for the expected treatment effect in an unspecified stratum. In this case, the more general null hypothesis under consideration ($\theta = \theta_0$) is regarding the underlying mean of the parameter across all strata, rather than a value that is assumed to be common to all strata. It would be possible to extend the **SCAS** method to incorporate a 'between stratum' component $\widehat{\tau}^2$ into the estimated variance of $S_j(\theta)$, using formulae based on the DerSimonian-Laird method ('**DL**').[25] However, the **DL** method has been found to have unsatisfactory performance (partly due to inadequate estimation of τ^2) and an alternative 'Hartung-Knapp-Sidik-Jonkman' method ('**HKSJ**')[26,27] based on the t-distribution has been recommended.[28,29] By adapting the formulae defining that method, a t-distribution asymptotic score method ('**TDAS**') may be devised. Full details of this new random effects method are given in Appendix S4.2.3 (available online).

4.2. Preliminary evaluation of stratified methods

As with the earlier examples for single stratum calculations, it is not possible to use any single example to draw conclusions on the relative performance of different methods. Selected methods are applied to a real meta-analysis dataset in Appendix B for software validation purposes.

There follows a brief evaluation of non-coverage probabilities estimated via simulations, comparing the stratified (fixed effects) **SCAS** method against the standard approximate normal ('**AN**') meta-analysis methods provided in the R package 'meta'.[30] Random effects methods are also shown, including **DL**, **HKSJ**, and the proposed new **TDAS** method.

Consistent with the evaluation of RNCP described earlier (but for a more restricted set of conditions), an examination of one-sided con-coverage rate is shown in Figure 4. Each plotted estimate of RNCP is based on 10,000 simulations, which is enough for a confidence interval width of approximately ± 0.003 for a nominal 0.025 significance level. This is generally sufficient for demonstrating the conditions under which the performance of some methods is very poor, but not precise enough to tease out more subtle differences between methods (such as **SCAS** and **MN**), or to rigorously demonstrate control of type I error within [0.0225, 0.0275]. Full details of the simulation study are given in Appendix S5 (available online).

Figure 4 displays the simulated type I error rates (right-sided non-coverage) for binomial RD for seven stratified methods evaluated at $p_1 = 0.2$, $p_2 = 0.1$ (with the individual stratum rates allowed to vary around those values), and $\alpha = 0.05$. The number of strata (k) ranges from 2 to 20, and the overall sample size $\sum_j N_j$ is $k \times 200$, with both equal and unequal treatment allocations. Two patterns of stratification are considered, one where all strata are of equal size ($N_j = 200$, $j = 1$ to k, 'pattern 1'), and an extreme case ('pattern 2') where one stratum is around 10 times the size of the others (e.g. with $k = 4$, $N_1 = 615$ and $N_j = 62$, $j = 2, 3, 4$). Finally, the amount of underlying heterogeneity is modelled using I^2 values (representing the proportion of variability due to heterogeneity[31]) of 0% (homogeneous) and 25% (modest heterogeneity). MH, IV and IVS weights are compared for the fixed effects methods, whereas random effects methods are restricted to IV weights (or IVS weights for **TDAS**).

The asymptotic normal method with IV weights has very poor coverage properties, which deteriorate with increasing number of strata. This reflects an inherent bias in the crude estimation of IV weights, as explained by Senn.[32] In contrast, the **AN** method with MH weights performs

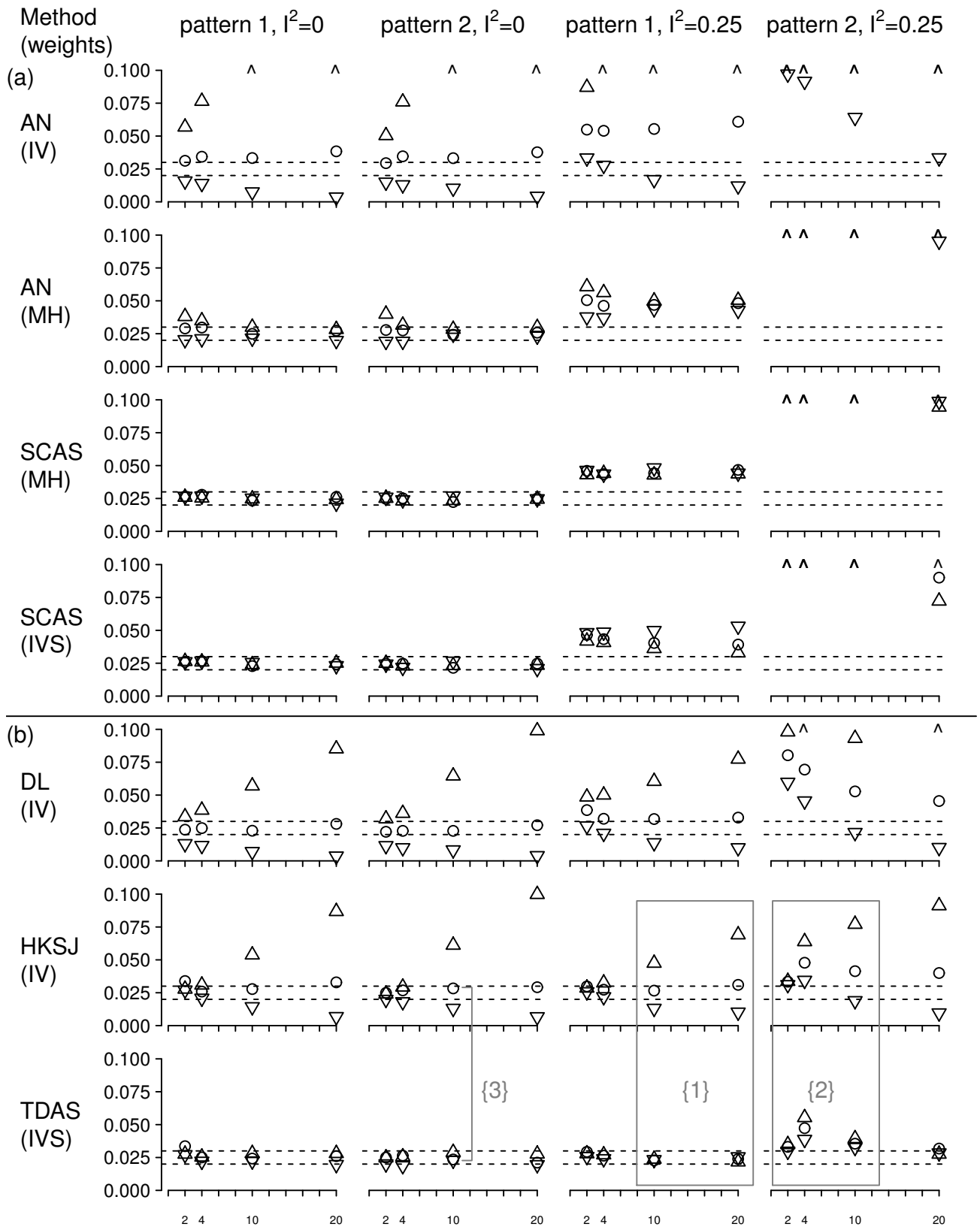

Figure 4. Simulated right-sided type I error rate for meta-analysis of binomial RD with different number of strata and sample size allocation (△ 3:1, ○ 1:1, ▽ 1:3), under four different sets of conditions. Pattern 1: equal-sized strata; Pattern 2: one stratum 10× larger than other strata; I^2=0: homogeneous treatment effects across strata; I^2=0.25: modest heterogeneity. True overall event rates $p_1 = 0.2$, $p_2 = 0.1$. Reference lines: $\alpha/2 \pm 0.2\alpha/2$. ^ indicates RNCP >0.1. (a): Fixed effects methods, (b): Random effects methods

surprisingly well with a large number of strata. For stratified **SCAS**, the coverage properties are superior to **AN** for any number of strata, and here the new IVS weights perform just as well as MH weights. A larger number of simulations would be required to confirm whether the stratified **SCAS** interval outperforms the stratified **MN** method.

Where stratum heterogeneity is present (see columns 3 and 4 of Figure 4), it comes as no surprise that the performance of the fixed effects methods deteriorates. (Nevertheless, the **SCAS**(MH) method remains valid for a test against the null hypothesis that $p_{1j} = p_{2j}$ for all j, as the presence of heterogeneity provides evidence against that null hypothesis.) For the random effects methods, the superiority of **HKSJ** over **DL** (as demonstrated by IntHout et al.[28]) is confirmed. However, the effect of the bias in IV weights is apparent again here, especially when treatment allocation is unequal. IVS weights for **TDAS** produce superior coverage under those conditions, and also with equal allocation when the number of strata is large (see region {1} of Figure 4).

Observations made by IntHout et al regarding situations with a small number of heterogeneous strata with diverse sizes apply also to **TDAS**. That is, under those conditions, type I error rates can be somewhat inflated (see region {2} of Figure 4). It appears that no method is completely satisfactory under such conditions.

Similar patterns were observed for all contrasts for both binomial and Poisson rates. However, for RR, a peculiar feature was observed with a very small number of strata. **TDAS** here tends too often to produce infinite-width intervals, and consequently the non-coverage probabilities are very low. This may reflect a weakness in the assumptions regarding the distribution of stratum estimates for RR in the derivation of the t-score. For plots of RNCP for the other contrasts, and further observations about the **TDAS** method in general, see Appendix S5 (available online).

In general, **TDAS** was the most consistent method in terms of proximity of RNCP to the nominal $\alpha/2$, under simulated conditions with and without heterogeneity (except for RR with very few strata). **SCAS** achieved this consistency under homogeneous conditions only.

It must be acknowledged that, due to the increased dimensionality when considering stratified designs, the evaluation carried out here has been necessarily limited. In particular, only a single parameter space point has been considered, although the variation of $\bar{p}$ and θ across strata means that sampling is taken across a fairly wide area of the parameter space. Secondly, this evaluation has used quite large stratum sizes, and the performance of these methods with smaller datasets may be of interest. Thirdly, conditions where the allocation ratio varies across strata have not been considered, and a wider range of stratum size patterns should perhaps also be considered for a thorough assessment of MH weights. Finally, as stated above, the restricted number of simulations due to time constraints means that subtle differences between some methods can be missed, as they are dwarfed by the very severe effect of the IV weighting scheme. For example, with 100,000 simulations (giving a confidence interval of ± 0.001) for the 10-stratum, pattern 1, homogeneous scenario with equal allocation, the estimated RNCP for **HKSJ** is slightly too high at 0.029, whereas that for **TDAS** is 0.025 (see region {3} of Figure 4).

On the basis of this limited evaluation, it appears that the **TDAS** method may have merit for meta-analysis under a wide range of conditions. It offers a substantial improvement over **HKSJ** with a large number of strata, or with unequal treatment allocation, although it seems that such analyses in practice are quite unusual.[33] With fewer than 10 strata, as long as stratum sizes are similar, **HKSJ** may be adequate with balanced treatment allocation. There may not be any satisfactory method if stratum sizes vary wildly, such as when a small meta-analysis is updated with the results of a large-scale confirmatory trial. Further work is required to further explore the performance of these methods.

5. CONCLUSION AND DISCUSSION

In summary, in the single stratum case, the **SCAS** methods achieve superior proximate equal-tailed coverage for all comparisons of binomial and Poisson rates, in all of the situations explored in this article. Of particular note is that non-coverage rates are improved even with rare events, and with small or large sample sizes. The skewness correction is particularly important for analysis of rate ratios, and in the Poisson case this correction has not previously been available. **MOVER-J** also performs reasonably well for RR with fairly large sample sizes, and **MOVER-B** is a convenient choice for all contrasts if prior information is to be incorporated in a Bayesian analysis. The simple asymptotic normal methods should always be avoided.

Interval location and interval width are often in direct competition, and the choice between them can be viewed as one of accuracy over precision. Much of the literature in this area discusses optimising interval width, without considering the effect on interval location. Nurminen[19] stated that equality of tail probabilities is a 'conventional optimality requirement', and it does seem likely that, for most users of two-sided confidence intervals, there is an implicit assumption that the tail probabilities are symmetrically distributed. If one seeks to minimise interval width, there is a risk of severely violating that assumption, and ending up with an interval that is precisely in the wrong place.

It is quite noticeable that for RR and OR, interval widths for all methods can be quite large when the number of events is very small, and the size of the denominators has little effect, or in the Poisson case, no effect at all. For Poisson RR this is quite natural, since increasing both denominators by the same factor is equivalent to selecting a different measurement unit for the exposure time (e.g. person-months instead of person-years). Such adjustments

should be expected to have no effect on the result because there is no material change to the data. For the binomial ratio contrasts, a similar feature is observed for rare events, when OR and binomial and Poisson RR are approximations of each other. This phenomenon suggests that the sample size required for such an analysis should be decided based on the number of observed events, not the total number of patients, similar to the approach used in survival analysis.

For stratified datasets, the performance of **SCAS** is also excellent for a fixed effects meta-analysis, i.e. when it can be assumed *a priori* that treatment effects are homogeneous across strata, or when the inferential aims of the analysis support the use of fixed effects.[34] If any heterogeneity is present, however, type I error rates are consistently inflated.

For a random effects meta-analysis, **TDAS** or **HKSJ** are recommended. The former is somewhat superior, but under balanced treatment allocation the difference can appear small in the context presented here, where much larger differences exist between some other methods. None of the random effects methods achieve satisfactory performance with few strata of diverse sizes. It has been argued that a small number of trials should not be combined in a meta-analysis at all if there is evidence of any heterogeneity.[35] Furthermore, for random effects methods, it may be difficult to justify a small number of strata as being a reasonable random sample from a population of strata, particularly if they correspond to studies or centers within a multi-center study.

The **TDAS** and **HKSJ** methods do not require a choice to be made in advance between a fixed or random effects approach (the performance of each method is largely unaffected by the presence of heterogeneity). However, if there is good reason to believe that the treatment effect is constant across strata, the stratified ('fixed effects') **SCAS** method may be preferred due to superior power (as discussed in Appendix S5.2, available online), especially if the number of strata is small. The **TDAS** method appears to achieve superior equal-tailed coverage compared to other available 'random effects' methods, particularly when sample sizes are unequal. Further work is required to assess in more detail the performance of these methods using different weighting strategies, under a wider range of conditions. To protect against undue influence by a small number of anomalous trials, a trimmed/Winsorised approach may be applied,[36] but it is perhaps rare that enough strata are available for this to be a viable option.

Note that the evaluation of meta-analysis methods in this paper (like others) is restricted to a particular model of heterogeneity (i.e. having normally distributed dispersion of RD, ln(RR) or ln(OR) between strata), which may or may not hold in practice. For example, if it is really the case that treatment effects are constant on the odds ratio scale, then the expected type I error of an analysis on the rate difference scale might behave quite differently.

The special case of stratified analysis of binomial data where $n_{ij} = 1$ for all i, j (i.e. matched pairs) has not been considered here. The stratified **MN** method has been found to be very conservative (too wide) in this situation.[37] The Tango score method,[38] designed for analysis of paired RD, is recognised as having good two-sided coverage[39] and location.[40] A corresponding asymptotic score method for paired RR is also available.[41] Based on a very cursory exploratory simulation study of the one-sided coverage properties of the stratified asymptotic score methods, assuming correlated event rates within strata, and with $n_{ij} = 1$, it appears that **TDAS** with MH weighting performs well for paired binomial RD and RR, and also for the corresponding paired Poisson comparisons with low expected event counts in a large number of strata. (**HKSJ** can be used for paired binomial RD only: for the other contrasts it is biased by the addition of 0.5 to zero cell counts.) The coverage properties under a wider range of conditions need further evaluation, but it seems that the **TDAS** method could be used for the analysis of paired rates, including Poisson rates, which are not catered for by existing methods. It would also be interesting to explore whether the one-sided performance of MOVER methods for paired RD[42] and RR[43] would be improved by the use of Jeffreys intervals.

Paired odds ratio, however, is another matter, and stratified **TDAS** does not appear to excel here (which may suggest problems also for stratified **TDAS** for OR when there are many sparsely-populated strata). As the best existing methods for the paired situation are based on transforming an interval for a single proportion,[40] it seems likely that either the **SCAS** or Jeffreys method for the single proportion would be good alternatives to the 'Transformed Clopper-Pearson mid-p' method. Both would have advantages of relative computational simplicity: the former has a closed-form solution, and the latter depends only on quantiles of a Beta distribution.

Until now, few of the above methods have been implemented in statistical software packages. An extension package ('ratesci'[44]) is now available for the free software R[45], containing the methods described in this paper, including **SCAS**, **MOVER-B** and **TDAS**. **HKSJ** is already available in the 'meta' package,[30] along with **AN** and **DL**. Of the above methods, SAS currently only provides **MN** and **AN** for binomial contrasts.[46] An online calculator for the single-stratum **SCAS** method can be found at `http://ssu.sheffield.ac.uk/ratesci/calc`.

REFERENCES

[1] Miettinen O, Nurminen M. Comparative analysis of two rates. *Statistics in Medicine* **1985**; 4(2):213–226, doi:10.1002/sim.4780040211.

[2] Laud PJ, Dane A. Confidence intervals for the difference between independent binomial proportions: comparison using a graphical approach and moving averages. *Pharmaceutical Statistics* **2014**; 13(5):294–308, doi:10.1002/pst.1631.

[3] Fagerland MW, Lydersen S, Laake P. Recommended confidence intervals for two independent binomial proportions. *Statistical Methods in Medical Research* **2015**; 24(2):224–254, doi: 10.1177/0962280211415469.

[4] Newcombe RG. Interval estimation for the difference between independent proportions: comparison of eleven methods. *Statistics in Medicine* **1998**; 17(8):873–890, doi:10.1002/(SICI)1097-0258(19980430)17:8$<$873::AID-SIM779$>$3.0.CO;2-I.

[5] Gart JJ, Nam Jm. Approximate interval estimation of the ratio of binomial parameters: A review and corrections for skewness. *Biometrics* **1988**; 44(2):323–338, doi:10.2307/2531848.

[6] Gart JJ, Nam Jm. Approximate interval estimation of the difference in binomial parameters: Correction for skewness and extension to multiple tables. *Biometrics* **1990**; 46(3):637–643, doi:10.2307/2532084.

[7] Newcombe RG. Measures of location for confidence intervals for proportions. *Communications in Statistics - Theory and Methods* **2011**; 40(10):1743–1767, doi:10.1080/03610921003646406.

[8] Newcombe RG, Nurminen MM. In defence of score intervals for proportions and their differences. *Communications in Statistics - Theory and Methods* **2011**; 40(7):1271–1282, doi:10.1080/03610920903576580.

[9] Brown LD, Cai TT, DasGupta A. Interval estimation for a binomial proportion. *Statistical Science* **2001**; 16(2):101–133, doi:10.1214/ss/1009213286.

[10] Agresti A, Coull BA. Approximate is better than "exact" for interval estimation of binomial proportions. *The American Statistician* **1998**; 52(2):119–126, doi:10.2307/2685469.

[11] Farrington CP, Manning G. Test statistics and sample size formulae for comparative binomial trials with null hypothesis of non-zero risk difference or non-unity relative risk. *Statistics in Medicine* **1990**; 9(12):1447–1454, doi:10.1002/sim.4780091208.

[12] Bartlett MS. Approximate confidence intervals. *Biometrika* **1953**; 40(1-2):12–19, doi:10.1093/biomet/40.1-2.12.

[13] Cornish EA, Fisher RA. Moments and cumulants in the specification of distributions. *Revue de l'Institut International de Statistique / Review of the International Statistical Institute* **1938**; 5(4):307–320, doi:10.2307/1400905.

[14] Donner A, Zou G. Closed-form confidence intervals for functions of the normal mean and standard deviation. *Statistical Methods in Medical Research* **2012**; 21(4):347–359, doi:10.1177/0962280210383082.

[15] Li HQ, Tang ML, Wong WK. Confidence intervals for ratio of two poisson rates using the method of variance estimates recovery. *Computational Statistics* **2014**; 29(3-4):869–889, doi:10.1007/s00180-013-0467-9.

[16] Cai TT. One-sided confidence intervals in discrete distributions. *Journal of Statistical Planning and Inference* **2005**; 131(1):63 – 88, doi:10.1016/j.jspi.2004.01.005.

[17] Brown LD, Cai TT, DasGupta A. Interval estimation in exponential families. *Statistica Sinica* **2003**; 13(1):19–49. Available at: http://www.jstor.org/stable/24307093 (Accessed 16.03.2017).

[18] Fagerland MW, Newcombe RG. Confidence intervals for odds ratio and relative risk based on the inverse hyperbolic sine transformation. *Statistics in Medicine* **2013**; 32(16):2823–2836, doi:10.1002/sim.5714.

[19] Nurminen M. Comparative analysis of two rates. some new approaches. PhD Thesis, University of Helsinki, Helsinki 1984. Available at: http://www.researchgate.net/publication/35217514 (Accessed 16.03.2017).

[20] Nurminen M, Mutanen P. Exact Bayesian analysis of two proportions. *Scandinavian Journal of Statistics* **1987**; 14(1):67–77, doi:10.2307/4616049.

[21] Rothman KJ, Greenland S, Lash TL. *Modern Epidemiology*. 3 edn., Lippincott Williams and Wilkins, 2008, pp. 244-249.

[22] Agresti A, Min Y. Frequentist performance of Bayesian confidence intervals for comparing proportions in 2 x 2 contingency tables. *Biometrics* **2005**; 61(2):515–523, doi:10.2307/3695972.

[23] Barker L, Cadwell BL. An analysis of eight 95 per cent confidence intervals for a ratio of Poisson parameters when events are rare. *Statistics in Medicine* **2008**; 27(20):4030–4037, doi:10.1002/sim.3234.

[24] Mehrotra DV, Railkar R. Minimum risk weights for comparing treatments in stratified binomial trials. *Statistics in Medicine* **2000**; 19(6):811–825, doi:10.1002/(SICI)1097-0258(20000330)19:6$<$811::AID-SIM390$>$3.0.CO;2-Z.

[25] DerSimonian R, Laird N. Meta-analysis in clinical trials. *Controlled Clinical Trials* **1986**; 7(3):177–188, doi:10.1016/0197-2456(86)90046-2.

[26] Hartung J, Knapp G. A refined method for the meta-analysis of controlled clinical trials with binary outcome. *Statistics in Medicine* **2001**; 20(24):3875–3889, doi:10.1002/sim.1009.

[27] Sidik K, Jonkman JN. A simple confidence interval for meta-analysis. *Statistics in Medicine* **2002**; 21(21):3153–3159, doi:10.1002/sim.1262.

[28] IntHout J, Ioannidis JPA, Borm GF. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. *BMC Med Res Methodol* **2014**; 14:25, doi:10.1186/1471-2288-14-25.

[29] Guolo A, Varin C. Random-effects meta-analysis: the number of studies matters. *Statistical Methods in Medical Research* **2015**; doi:10.1177/0962280215583568.

[30] Schwarzer G. meta: An R package for meta-analysis. *R News* **2007**; 7(3):40–45.

[31] Higgins JPT, Thompson SG. Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine* **2002**; 21(11):1539–1558, doi:10.1002/sim.1186.

[32] Senn S. *Statistical Issues in Drug Development*. 2 edn., John Wiley & Sons, Ltd., 2007, p264.

[33] Davey J, Turner RM, Clarke MJ, Higgins JP. Characteristics of meta-analyses and their component studies in the Cochrane database of systematic reviews: a cross-sectional, descriptive analysis. *BMC Medical Research Methodology* **2011**; 11(1):160, doi:10.1186/1471-2288-11-160.

[34] Hedges LV, Vevea JL. Fixed- and random-effects models in meta-analysis. *Psychological Methods* **1998**; 3(4):486–504, doi:10.1037/1082-989X.3.4.486.

[35] Gonnermann A, Framke T, Großhennig A, Koch A. No solution yet for combining two independent studies in the presence of heterogeneity. *Statistics in Medicine* **2015**; 34(16):2476–2480, doi:10.1002/sim.6473.

[36] Emerson JD, Hoaglin DC, Mosteller F. Simple robust procedures for combining risk differences in sets of 2 x 2 tables. *Statistics in Medicine* **1996**; 15(14):1465–1488, doi:10.1002/sim.4780151402.

[37] Klingenberg B. A new and improved confidence interval for the Mantel–Haenszel risk difference. *Statistics in Medicine* **2014**; 33(17):2968–2983, doi:10.1002/sim.6122.

[38] Tango T. Equivalence test and confidence interval for the difference in proportions for the paired-sample design. *Statistics in Medicine* **1998**; 17(8):891–908, doi:10.1002/(SICI)1097-0258(19980430)17:8<891::AID-SIM780>3.0.CO;2-B.

[39] Newcombe RG. Author's reply. *Statistics in Medicine* **1999**; 18(24):3513–3513, doi:10.1002/(SICI)1097-0258(19991230)18:24<3513::AID-SIM304>3.0.CO;2-1.

[40] Fagerland MW, Lydersen S, Laake P. Recommended tests and confidence intervals for paired binomial proportions. *Statistics in Medicine* **2014**; 33(16):2850–2875, doi:10.1002/sim.6148.

[41] Tang NS, Tang ML, Chan ISF. On tests of equivalence via non-unity relative risk for matched-pair design. *Statistics in Medicine* **2003**; 22(8):1217–1233, doi:10.1002/sim.1213.

[42] Newcombe RG. Improved confidence intervals for the difference between binomial proportions based on paired data. *Statistics in Medicine* **1998**; 17(22):2635–2650, doi:10.1002/(SICI)1097-0258(19981130)17:22<2635::AID-SIM954>3.0.CO;2-C.

[43] Tang ML, Li HQ, Tang NS. Confidence interval construction for proportion ratio in paired studies based on hybrid method. *Statistical Methods in Medical Research* **2012**; 21(4):361–378, doi:10.1177/0962280210384714.

[44] Laud P. *ratesci: Confidence Intervals for Comparisons of Binomial or Poisson Rates* 2016. R package version 0.1-0. http://cran.r-project.org/web/packages/ratesci/.

[45] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria 2016.

[46] SAS Institute Inc. *SAS/STAT® 14.2 User's Guide*. SAS Institute Inc., Cary, NC 2016.

[47] Graham PL, Mengersen K, Morton AP. Confidence limits for the ratio of two rates based on likelihood scores: non-iterative method. *Statistics in Medicine* **2003**; 22(12):2071–2083, doi:10.1002/sim.1405.

[48] Wilson EB. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association* **1927**; 22(158):209–212, doi:10.1080/01621459.1927.10502953.

APPENDIX A DEFINITIONS FOR SKEWNESS-CORRECTED SCORE METHODS

Within the unified definition of the score method shown in Section 2.1, the form of $S(\theta)$, and the estimation of its variance and skewness, depend on the situation as specified below. These estimates rely on obtaining $\widetilde{p}_2$, the restricted MLE of p_2, from which $\widetilde{p}_1$ follows as $\widetilde{p}_2 + \theta$ (for RD), $\widetilde{p}_2\theta$ (for RR), or $\widetilde{p}_2\theta/(1 + \widetilde{p}_2(\theta - 1))$ (for OR). The formulae for $S(\theta)$, $\widetilde{p}_2$ and $\widetilde{V}$ are essentially reproduced from Miettinen and Nurminen.[1] The components that are added here are the skewness estimates. These are derived for each of the score statistics, using an estimate of the third central moment $\widetilde{\mu}_3$ for each case. This is obtained by treating θ, $\widetilde{p}_1$ and $\widetilde{p}_2$ as constants, and using the fact that for $c\widehat{p}$, a constant multiple of the estimator of a single binomial proportion, $\mu_3(c\widehat{p})$ is $c^3p(1-p)(1-2p)/n^2$, while for a Poisson rate it is c^3p/n^2. For binomial RD and RR, these skewness corrections are found to be theoretically or empirically equivalent to those derived by Gart and Nam.[5,6] I believe the skewness corrections applied to the Poisson case and the odds ratio are novel developments.

A.1 Rate Difference

The contrast function for RD is $S(\theta) = \widehat{p}_1 - \widehat{p}_2 - \theta$, regardless of whether binomial or Poisson rates are involved. The required estimated central moments for the binomial case are:

$$\widetilde{V} = [\widetilde{p}_1(1-\widetilde{p}_1)/n_1 + \widetilde{p}_2(1-\widetilde{p}_2)/n_2]N/(N-1)\ ^{\dagger}$$

$$\widetilde{\mu}_3 = [\widetilde{p}_1(1-\widetilde{p}_1)(1-2\widetilde{p}_1)/n_1^2] - [\widetilde{p}_2(1-\widetilde{p}_2)(1-2\widetilde{p}_2)/n_2^2]$$

where $\widetilde{p}_2$ is obtained by solving the cubic equation $A\widetilde{p}_2^3 + B\widetilde{p}_2^2 + C\widetilde{p}_2 + D = 0$, with:

$A = N$;
$B = (n_1 + 2n_2)\theta - N - X$;
$C = [n_2\theta - N - 2X_2]\theta + X$;
$D = X_2\theta(1-\theta)$;
$X = X_1 + X_2$ and $N = n_1 + n_2$.

This equation has a closed-form solution given by $\widetilde{p}_2 = 2u\cos(w) - B/(3A)$,

where $w = \frac{1}{3}\left[\pi + \cos^{-1}\frac{v}{u^3}\right]$; $v = \frac{B^3}{(3A)^3} - \frac{BC}{6A^2} + \frac{D}{2A}$; and $u = sign(v)\left[\frac{B^2}{(3A)^2} - \frac{C}{3A}\right]^{1/2}$.

Note that it is necessary to truncate the input of the cos^{-1} function in w to the range $[0, 1]$ to avoid errors from 'negative zero' values due to floating-point arithmetic used in computer software. Also, if $u = 0$, $\widetilde{p}_2 = -B/(3A)$.

For the Poisson case, the central moments of $S(\theta)$ are estimated using:

$$\widetilde{V} = \widetilde{p}_1/n_1 + \widetilde{p}_2/n_2$$

$$\widetilde{\mu}_3 = \widetilde{p}_1/n_1^2 - \widetilde{p}_2/n_2^2$$

where $\widetilde{p}_2$ is the solution (in $[0, 1]$) of the quadratic $A\widetilde{p}_2^2 + B\widetilde{p}_2 + C = 0$, that is:

$$\widetilde{p}_2 = [-B + (B^2 - 4AC)^{1/2}]/(2A),$$

where $A = N$, $B = N\theta - X$ and $C = -X_2\theta$.

A.2 Rate Ratio

For RR, the test statistic is based on the Fieller-type contrast $S(\theta) = \widehat{p}_1 - \widehat{p}_2\theta$. The central moments for the binomial case are estimated with:

$$\widetilde{V} = [\widetilde{p}_1(1-\widetilde{p}_1)/n_1 + \theta^2\widetilde{p}_2(1-\widetilde{p}_2)/n_2]N/(N-1)$$

$$\widetilde{\mu}_3 = \widetilde{p}_1(1-\widetilde{p}_1)(1-2\widetilde{p}_1)/n_1^2 - \theta^3[\widetilde{p}_2(1-\widetilde{p}_2)(1-2\widetilde{p}_2)/n_2^2]$$

where $\widetilde{p}_2 = [-B - (B^2 - 4AC)^{1/2}]/(2A)$,

$A = N\theta$, $B = -[n_1\theta + X_1 + n_2 + X_2\theta]$, and $C = X$.

Note that if $\theta = 0$, then $A = 0$, so the quadratic equation in $\widetilde{p}_2$ becomes a linear equation $B\widetilde{p}_2 + C = 0$, and therefore $\widetilde{p}_2 = X/(X_1 + n_2)$.

In the case of Poisson RR, Graham et al[47] pointed out that since $\widetilde{p}_2$ is the solution of a linear equation, the asymptotic score (**MN**) interval has a closed form solution, being the roots of a quadratic. However, to incorporate the skewness correction requires iteration as before, using:

$$\widetilde{V} = \widetilde{p}_1/n_1 + \theta^2\widetilde{p}_2/n_2$$

$$\widetilde{\mu}_3 = \widetilde{p}_1/n_1^2 - \theta^3\widetilde{p}_2/n_2^2$$

where $\widetilde{p}_2 = X/(n_1\theta + n_2)$

[†]Note the inclusion of the factor $N/(N-1)$ in $\widetilde{V}$, omitted by Gart and Nam, which results in an unbiased variance estimate, and improves coverage properties for smaller sample sizes.

A.3 Odds Ratio

For OR, the situation is slightly different: the score function $S(\theta)$ is defined in terms of $\widetilde{p}_1$ and $\widetilde{p}_2$ instead of θ, but these are still solved as functions of θ, so in practice the algorithm can proceed in the same way as for RD or RR. In this case:

$$S(\theta) = \frac{\widehat{p}_1 - \widetilde{p}_1}{\widetilde{p}_1(1-\widetilde{p}_1)} - \frac{\widehat{p}_2 - \widetilde{p}_2}{\widetilde{p}_2(1-\widetilde{p}_2)},$$

or equivalently

$$S(\theta) = (X_1 - n_1\widetilde{p}_1)\left[\frac{1}{n_1\widetilde{p}_1(1-\widetilde{p}_1)} + \frac{1}{n_2\widetilde{p}_2(1-\widetilde{p}_2)}\right]$$

$$\widetilde{V} = [1/(n_1\widetilde{p}_1(1-\widetilde{p}_1)) + 1/(n_2\widetilde{p}_2(1-\widetilde{p}_2))]N/(N-1)$$

$$\widetilde{\mu}_3 = (1-2\widetilde{p}_1)/(n_1\widetilde{p}_1(1-\widetilde{p}_1))^2 - (1-2\widetilde{p}_2)/(n_2\widetilde{p}_2(1-\widetilde{p}_2))^2$$

where $\widetilde{p}_2 = [-B + (B^2 - 4AC)^{1/2}]/(2A)$, with

$A = n_2(\theta - 1)$, $B = n_1\theta + n_2 - X(\theta - 1)$, $C = -X$.

Note that if $\theta = 1$, then $A = 0$, so
$\widetilde{p}_2 = -C/B = X/N$.

A.4 Single binomial or Poisson rate

It is also possible to apply the skewness correction to the Wilson score method[48] for a single proportion ($\theta = p$), and the corresponding Poisson interval, using $S(\theta) = \widehat{p} - \theta$. In this case, restricted maximum likelihood estimation of p for a given θ is unnecessary (since $p = \theta$): V and μ are calculated directly as $V = \theta(1-\theta)/n$ and $\mu_3 = \theta(1-\theta)(1-2\theta)/n^2$ for a binomial proportion, and $V = \theta/n$ and $\mu_3 = \theta/n^2$ for a Poisson rate. In both cases, closed-form solutions can be calculated:

For the binomial proportion, the limits are found using $[-B \pm (B^2 - 4AC)^{1/2}]/(2A)$, where

$A = E^2 + z^2/n$, $B = 2DE - z^2/n$, $C = D^2$,
$D = (z^2 - 1)/(6n) - x/n$, $E = 1 - (z^2 - 1)/(3n)$.

The Poisson confidence limits are calculated similarly, but with $A = 1$, $B = 2D - z^2/n$, $C = D^2$, and $D = (z^2 - 1)/(6n) - x/n$.

When $x = 0$ the lower limit is corrected to 0, and in the binomial case the upper limit is 1 when $x = n$.

A.5 Skewness-corrected hypothesis test

For a hypothesis test against $\theta = \theta_0$, the Z-score can be calculated from Equation (3) as $[-B + (B^2 - 4AC)^{1/2}]/(2A)$, where
$A = \widetilde{\mu}_3/(6\widetilde{V})$, $B = \widetilde{V}^{1/2}$ and $C = -[S(\theta_0) + \widetilde{\mu}_3/(6\widetilde{V})]$, with $\widetilde{\mu}_3$ and $\widetilde{V}$ evaluated at $\theta = \theta_0$.

In the stratified case,
$A = \sum_j (w_j/W)^3\widetilde{\mu}_{3j}/(6\widetilde{V}_\cdot)$, $B = \widetilde{V}_\cdot^{1/2}$ and
$C = -\sum_j[(w_j/W)S_j(\theta_0) + (w_j/W)^3\widetilde{\mu}_{3j}/(6\widetilde{V}_\cdot)]$.

APPENDIX B EXAMPLES

Hartung and Knapp[26] presented an example of a meta-analysis of placebo-controlled trials of cisapride for the treatment of non-ulcer dyspepsia. The dataset is provided within the R package 'meta', so it is not reproduced here.

Example confidence intervals for the single-stratum case are shown in Table 2, using each of the methods described in Section 2. These examples use the results of two trials selected from the cisapride data, together with a more extreme example used by Newcombe,[4] comparing the proportion of patients experiencing respiratory adverse effects following fungicidal treatment with terbinafine versus placebo. The same data are used to produce artificial examples for the Poisson distribution methods (as if the 'success' outcome were switched to an adverse event, and assuming that the total follow-up period was proportional to the number of patients in each group).

These examples are presented for illustration and software validation, but note that it is not possible to draw general inferences about the relative performance of different methods from any single example. Readers who are tempted to choose the **MN** method for RR based on the relative interval widths shown here should refer back to Figure 2.

The full cisapride meta-analysis is presented in Table 3, showing point estimates and confidence intervals for the estimated overall RD, RR and OR using various methods. This includes the fixed effects approximate normal (**AN**) method with both IV and MH weights, and the **SCAS** intervals using MH and IVS weights. Random effects intervals are shown, including **DL** and **HKSJ**, which rely on IV weights, and **TDAS** with IVS weights.

For all three contrasts, significant heterogeneity ($I^2 = 73\%$) is confirmed using the test described in Appendix S4.2.1, and the random effects confidence intervals are accordingly substantially wider than the fixed effects intervals.

Table 2. Example 95% Confidence Intervals

	Success rate: 12/16 (active) vs 1/16 (placebo) (Milo 1984)				
	Binomial RD	Poisson RD	Binomial RR	Poisson RR	Binomial OR
SCAS	(0.386, 0.878)	(0.285, 1.221)	(2.648, 204.300)	(2.161, 221.836)	(5.586, 1025.364)
MN	(0.375, 0.863)	(0.299, 1.255)	(2.487, 69.950)	(2.002, 71.937)	(5.144, 349.002)
MOVER-J	(0.373, 0.849)	(0.274, 1.201)	(2.692, 109.329)	(2.214, 115.261)	(5.907, 521.865)
AN	(0.444, 0.931)	(0.246, 1.129)	(1.762, 81.745)	(1.560, 92.287)	(4.426, 457.475)
	19/29 vs 22/30 (Kellow 1995)				
	Binomial RD	Poisson RD	Binomial RR	Poisson RR	Binomial OR
SCAS	(-0.312, 0.160)	(-0.519, 0.361)	(0.613, 1.271)	(0.480, 1.653)	(0.220, 2.126)
MN	(-0.309, 0.158)	(-0.524, 0.365)	(0.615, 1.270)	(0.488, 1.635)	(0.230, 2.081)
MOVER-J	(-0.301, 0.153)	(-0.511, 0.355)	(0.623, 1.257)	(0.482, 1.646)	(0.222, 2.096)
AN	(-0.313, 0.156)	(-0.503, 0.347)	(0.635, 1.256)	(0.484, 1.651)	(0.227, 2.105)
	Adverse event rate: 5/56 vs 0/29 (Goodfield 1992)				
	Binomial RD	Poisson RD	Binomial RR	Poisson RR	Binomial OR
SCAS	(-0.019, 0.187)	(-0.023, 0.197)	(0.770, ∞)	(0.726, ∞)	(0.755, ∞)
MN	(-0.033, 0.193)	(-0.043, 0.209)	(0.717, ∞)	(0.674, ∞)	(0.696, ∞)
MOVERJ	(-0.010, 0.177)	(-0.013, 0.188)	(0.851, 5473)	(0.804, 5460)	(0.839, 6030)
AN	(0.015, 0.164)	(0.011, 0.168)	(0.000, ∞)	(0.000, ∞)	(0.000, ∞)

Table 3. Stratified estimates and 95% Confidence Intervals for cisapride meta-analysis

	Method	Weights	RD	RR	OR
Fixed effects:	**AN**	IV	0.341 (0.282, 0.400)	1.56 (1.38, 1.76)	3.42 (2.54, 4.61)
	AN	MH	0.309 (0.249, 0.369)	1.76 (1.54, 2.00)	3.64 (2.73, 4.85)
	SCAS	MH	0.309 (0.246, 0.370)	1.76 (1.55, 2.00)	3.87 (2.88, 5.22)
	SCAS	IVS	0.308 (0.244, 0.370)	1.75 (1.55, 2.00)	3.91 (2.91, 5.28)
Random effects:	**DL**	IV	0.338 (0.213, 0.463)	1.75 (1.37, 2.22)	4.14 (2.34, 7.32)
	HKSJ	IV	0.338 (0.203, 0.473)	1.75 (1.31, 2.32)	4.14 (2.22, 7.73)
	TDAS	IVS	0.329 (0.193, 0.465)	1.83 (1.39, 2.58)	4.33 (2.32, 9.04)